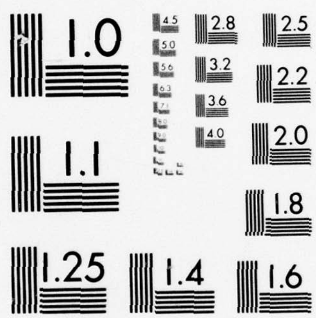
AD-A053 499 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF INDUSTRIA--ETC F/G 12/2
APPROXIMATION TECHNIQUES IN THE SOLUTION OF QUEUEING PROBLEMS.(U)
MAR 78 U N BHAT, M J FISCHER, M SHALABY N00014-75-C-0517
UNCLASSIFIED IEOR-76017 NL

1 OF 1
AD
A053499

END
DATE
FILMED
6-78
DDC




MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

6

B.S.

# SOUTHERN METHODIST UNIVERSITY





DEPARTMENT OF OPERATIONS RESEARCH AND ENGINEERING MANAGEMENT

SCHOOL OF ENGINEERING AND APPLIED SCIENCE

DALLAS, TEXAS 75275

# APPROXIMATION TECHNIQUES IN THE SOLUTION OF QUEUEING PROBLEMS

by

U. Narayan Bhat*
Martin J. Fischer**
Mohamed Shalaby*

* Department of Operations Research and Engineering Management
School of Engineering and Applied Science
Southern Methodist University
Dallas, Texas 75275

** Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, Virginia





Revised March 1978

* Research work supported by the ONR/DCEC Contract N00014-75-C-0597, NR042-324. Reproduction in whole or in part is permitted for any purpose of the United States Government.

## ABSTRACT

In the study of complex queueing systems analysis techniques aimed at providing exact solutions become ineffective. Approximation techniques provide an attractive alternative in such cases. This paper gives an overview of different types of approximation techniques available in the literature and points out their relative merits. Also the need for proper validation procedures of approximation techniques is emphasized.

Queueing theory has passed through several stages in its growth. The first three decades of this century consisted of pioneering work in its foundations. Major analysis techniques for the investigation into the behavior of Markovian systems were developed during the next two decades. The nineteen hundred and fifties saw investigations extended into problems related to non-Markovian systems. This trend continued well into the middle of the sixties. Until then, queueing theory, having been developed by mathematicians, probabilists and statisticians, had grown with minimal interaction with applications. During the past ten years, the trend has been more towards applications and making queueing theory results applicable. The two major areas receiving maximum attention during this period are optimization problems in queues and approximation technqiues in the solution of queueing problems.

As the complexity of the systems being considered by applied scientists increases, finding effective solution techniques leading to exact solutions is becoming a difficult task. Approximation techniques provide an attractive alternative in such cases. Over the years several types of approximation techniques have been developed for the solution of queueing problems. It is our intention here to provide an overview of these techniques and discuss their relative merits.

Three different stages may be identified in the modeling and analysis of a queueing system. At the first stage a suitable mathematical model for the system is developed. The second stage concerns the identification of and investigation into the basic process underlying the model. At the third stage numerical results are obtained from the analysis of the process. Noting that an approximation can

be introduced at any one of these stages, we may identify three major categories of approximation techniques: (1) system approximation, (2) process approximation, and (3) numerical approximation.

In the following sections we shall discuss different techniques used in approximations for the solutions of queueing problems based on the above categorization. Since justifying approximate results is an integral part of the process, techniques for validating approximations are also discussed. Finally comments are made about future prospects in this direction.

It should be pointed out, however, that the objective of the paper is not the categorization, but understanding different types of approximating procedures. As will be clear later, even though the approximation is initiated at a certain stage, the net result is to impact the system at all stages of analysis. Consequently the distinction between the techniques sometimes becomes unclear.

For purposes of convenience we use Kendall notation suitably modified to include finite capacity. For instance, A/B/C/D represents a system in which the symbols A, B, C, D stand for the inter-arrival time distribution, service time distribution, number of servers and system capacity respectively. When dealing with systems with no limitations on capacity, D is dropped. Also the time dependence of an element is indicated by writing it as a function of t.

In compiling a bibliography, the intention of the authors has been to include a representative list of references. We have also tried to be exhaustive so as to make it useful to the reader. All omissions of significant papers are inadvertent rather than intentional.

System Approximation

A system approximation is mainly a simplification of the system under study such that the behavior of the new system is strongly related to the original system. The four main elements in a queueing system are the arrival process, the queue discipline, the service process, and the system structure. These elements are described by their properties or attributes. Also, due to the complexity of some applications such as networks of queues we need to add a set of relations that hold among these elements, which are the results of various assumptions. Hence, a system simplification may be characterized either as simplifying the system elements or relaxing the relational assumptions.

Simplification of system elements is at the heart of the practice of queueing theory. Many times, results may not be available for the exact representation of the system element model (such as the distribution for inter-arrival time or service time). Then, the best available model is used to arrive at the best approximate result. The predominant use of the exponential distribution in practice is due to this approximating process. In an attempt to incorporate more general inter-arrival time and service time distributions, Erlangian distribution and Erlangian mixtures have been extensively used. In this regard the papers by Luchak (1956), Wishart (1959), Kotiah et al. (1969), Schassberger (1970) are significant. The first three of the above papers supply the practicality of the approach, whereas the last paper provides the theoretical basis for the procedure.

A common technique in system approximation is to use a simpler system either to derive an approximate measure of performance or suitable gounds for them. For instance Maaloe (1973) uses simple relations existing between mean waiting times of M/M/1 and M/M/s systems to provide an approximation value of the mean waiting time

in an $M/E_k/s$ system. Gross (1975) examines the effect of using an M/M/s model to approximate a G/G/s model. His results indicate that when estimating mean value measures of congestion, the sensitivity to the exponential assumption is more pronounced whereas it is not as pronounced for cost optimization models. Chandy et.al. (1975) study a queueing network with a direct application of Norton's theorem (Chandy et.al, 1975b) which implies that the properties of a subsystem in a queueing network can be obtained by replacing all queues that are not of interest by a single queue with equivalent load characteristics (also see Sauer and Chandy (1975), Chandy, Herzog and Woo (1975)). Another approach in the treatment of queueing networks occurring in computer systems is that of Avi-Itzhak and Heyman (1973). First, exact results are obtained for a closed-system model in terms of cycle times and server utilization. These results are then used to develop approximate results for an open system model. For other examples of the use of simpler systems see Ghosal (1970) and Rosenshine (1976).

Nonstationarity in the arrival process can also be effectively handled through approximations. Moore (1975) provides methods for partitioning the time axis into intervals with stationary characteristics and approximates a M(t)/G/1 queue by an M/G/1 queue during these periods.

Using simpler systems, upper and lower bounds for system performance measures have been derived in several cases. Brosh (1969) derives mean total time spent by a customer in a priority queueing system by essentially changing the priority level of the customer so as to provide a worse and a better case. Brumelle (1971) obtains bounds for mean waiting time in a G/G/s system by constructing two single server systems; one of them uses a share of the original load to give an upper bound and the second uses a service rate s times faster than

the original one to give a lower bound. A further improvement on the upper bound for mean waiting time in the system G/M/s is obtained by Brumelle (1973) by the waiting time in an associated G/M/1 queue. Yu (1974) bounds a multi-server queue with recurrent input and Erlang service times by a simple $G/E_k/1$ queue. Kotiah (1977) uses a linear programming technique to provide bounds in Markovian systems.

Simpler bounding systems can be obtained by modifications to the queue discipline. Under low traffic situations Bloomfield and Cox (1972) obtain lower bounds for mean waiting time by ignoring the waiting times of customers other than the one being considered. In the context of a traffic queue at a signalized road intersection Bhat and Prabhu (1975) obtain upper and lower bounds by sweeping the traffic arriving during a green period to the right and left extremities of the period [also see, Bhat, Wheeler and Fischer (1974)].

Replacing a general distribution by one that has the same moments is an appealing approach. Kuczura (1973) approximates the overflow process of an M/M/s/s system by an interrupted Poisson process which is alternately turned on and off for exponentially distributed lengths of time. The approximation is obtained by matching the first two or three moments of the two processes. To study the mean waiting time of a G/G/1 queue, Marchal and Harris (1976) use an $E_k/E_\ell/1$ queue and match the first four moments of the random variable representing the difference (service time - interarrival time).

A problem of great interest in telephone work relates to predicting the blocking probability of an overflow stream of traffic in a group of channels operating as a loss system. An approximation widely used is the equivalent random method which replaces the system under consideration by an equivalent loss system with a Poisson input. For details of this method see Wilkinson (1955), Cooper(1972) and Holtzman (1975).

There are queueing systems in which more than one class of customers share the resources. A relatively simple procedure to derive the performance measures of such systems is to consider the two classes separately and improve the accuracy of approximation by successively using the most recent results for one class in the derivation of results for the other (see Bhat and Raju (1977)).

For approximating more complex systems, many of these different characteristics could be used at different stages. Some examples of such efforts may be found in papers such as Leibowitz (1961), Halfin (1975), Willemain (1974), and Rosenshine and Chandra (1975).

Many system approximations are heuristic in nature. The quality of such procedures depend very much on intuition and creativity. Justification for the use of heuristic methods is not because they are analytically sound, but because experimentation has proved that they are useful in practice. The basic approach is to observe the system, relate it to some other system with known behavior and then make an educated guess about the behavior of the original system. For instance, Cosmetatos (1974) derives approximate formulae for the steady state queue size and waiting time distribution in the system GI/M/s, by observing the similarity of the mean waiting time curves drawn against the coefficient of variation of the inter-arrival time distribution, when the traffic intensity is kept constant for different number of servers. By this procedure he obtains approximate results that are within 5% of the actual value. Bhat and Fischer (1976) have derived approximate results such as blocking probability and waiting time in a two class heterogeneous multiserver system with Poisson arrivals, in which one class acts as a loss system but the second acts as a delay system. A key to this procedure is the observation

that the probability of blocking is relatively insensitive to the ratio of the service rates of each class, which allows them to assume equal service rates. Conolly (1975) considers Poisson queues belonging to the class of generalized birth and death process as essentially renewal models with "effective" inter-arrival and service times (actual intervals may be dependent on queue size).

Nozaki and Ross (1976) provide an approximation for mean waiting time in a multi-server queue M/G/s, by assuming the equilibrium distribution form for the remaining service time of customers in service at the time of arrival. The expression involves the distribution of the number of busy servers, for which an approximate formula similar to the exact distribution in the queue M/M/s, is derived.

Given above are only some examples of the use of heuristic approaches in approximations. To some extent all approximations can be considered to have some heuristic elements in it; but in system approximations they are in abundance.

Process Approximation

Representation of a mathematical model follows the identification of the system model. Many times, the basic process underlying the mathematical model is so complex that a direct analysis does not become worthwhile for the situation. One alternative would be to simplify the system model itself as described above. The second alternative is to identify simpler process whose analysis is either known or can be derived, that has properties similar to the basic process. Diffusion approximation, fluid approximation, and the use of asymptotic or limiting results are examples of such procedures. System approximation techniques described in the pervious section can also be

looked upon as a form of process approximation, when the availability of a simpler underlying process is the motivation for such an effort. System approximation techniques suggested by Moore (1975) and Bhat and Prabhu (1975) are examples of such situations.

Fluid approximation, as suggested by Newell (1971) is mostly an engineering approach. It starts with some crude and naive estimates and relationships between system elements and improvements are made in them as the analysis proceeds. The essential concept is to consider the arrival and departure processes in the system as fluid flowing in and out of a reservoir. Because of its deterministic nature when the output rate (service rate) is in excess of the input rate (arrival rate) the fluid approximation results in an empty queue. In view of this a proper setting for its application would be a short term analysis of a queue or the behavior of an over-saturated queue (when the arrival rate exceeds the service rate). Also the particular significance of its usage would be when the arrival and service rates are time dependent. Then if $A(t)$ and $D(t)$ are the arrival and departure processes, with rates $\lambda(t) = dA(t)/dt$ and $\mu(t) = dD(t)/dt$ respectively, an approximate expression for the queue length $Q(t)$ at time $t$ can be given as

$$\begin{aligned} Q(t) &\simeq Q(0) + A(t) - D(t) \\ &= Q(0) + \int_0^t \lambda(\tau)d\tau - \int_0^t \mu(\tau)d\tau. \end{aligned}$$

A stochastic analogue of the fluid approximation is the diffusion approximation. In this procedure we replace a queueing process with jump transitions or with continuous and jump transitions by a continuous process which reflects the main characteristics

of the original process. Diffusion processes are governed by stochastic differential equations incorporating the infinitesimal mean and variance of the process. Let

$$E\{Q(t+\tau) - Q(t)\} = \int_t^{t+\tau} [\lambda(x) - \mu(x)]dx$$

$$\simeq [\lambda(t) - \mu(t)]\tau$$

where the arrival and departure rates $\lambda(t)$ and $\mu(t)$ are considered to be nearly constant over times as compared to $\tau$. The quantity $\lambda(t) - \mu(t)$ is known as the infinitesimal mean of the process at time t. Also, let

$$\sigma^2(t) = \text{Var}\ \{Q(t+\tau) - Q(t)\}/\tau$$

be the infinitesimal variance of the process. Denoting the distribution of the process $Q(t)$ by $f(x,t)$ (note that, $Q(t)$ is considered to be a continuous process in this approximation), under this approximation, the function $f(x,t)$ is assumed to satisfy the Fokker-Planck equation

$$\frac{\partial f(x,t)}{\partial t} = -[\lambda(t) - \mu(t)]\ \frac{\partial f(x,t)}{\partial x} + \frac{\sigma^2(t)}{2}\ \frac{\partial^2(x,t)}{\partial x^2}\ .$$

Diffusion approximation is usually related to heavy traffic (service rate close to the arrival rate) since we need the time variable to be large as compared to intervals between transitions. Under heavy traffic idle periods occur very infrequently and therefore one could use the zero state as a reflecting barrier of the process without degrading the approximation much further (note that a diffusion process can drift towards states below 0, whereas a queueing process remains on the non-negative side of the axis).

The equilibrium distribution of the resulting process can be derived in most cases as $f(x) = \lim_{t\to\infty} f(x,t)$. If necessary it can be discretized by integrating it over the unit interval $n < x \leq n + 1$, or $n - .5 < x \leq n + .5$.

Gaver's (1968) analysis of the virtual waiting time of an M/G/1 queue is one of the initial effects in using diffusion approximation for queueing systems. In this case the infinitesimal mean and variance for the process are $\lambda E(S)-1$ and $\lambda E(S^2)$, respectively where s is the service time [also, see Gaver (1968)]. Newell (1968) gives an extensive treatment of a time dependent arrival process using the Fokker-Planck equation. Heyman (1974) has extended Gaver's (1968) results to study the busy period of the queue M/G/1. The transient behavior of the G/G/1 queue has been approximated by Heyman (1975), which has been extended to the G/G/k system by Halachmi and Franta (1976) by similar approximation techniques. Denoting the inter-arrival time by A and service time by S, for G/G/1, the infinitesimal mean and variance are

$$\lambda(t) - \mu(t) = \{\frac{1}{E(A)} - \frac{1}{E(S)}\}\ t$$

$$\sigma^2(t) = \{\frac{Var(A)}{[E(A)]^3} + \frac{Var(S)}{[E(S)]^3}\}\ t\ .$$

For the queue G/G/k, these take the form

$$\lambda(t) - \mu(t) = \{\frac{1}{E(A)} - \frac{\min(x,k)}{E(S)}\}\ t$$

$$\sigma^2(t) = \{\frac{Var(A)}{[E(A)]^3} + \frac{\min(x,k)Var(S)}{[E(S)]^3}\}\ t$$

where x is the state of the system.

Some of the other applications of diffusion approximation in queues can be found in Newell (1975), who provides a general setting for the analysis of the behavior of a sequence of servers in series

with finite storage in between. Other references that suggest and elaborate earlier applications can be found in Kimura (1964), Newell (1965), Cox and Miller (1965) and Feller (1966).

Diffusion approximation has also been successfully employed in the analysis of queueing networks. Appropriate references in this area are Kobayashi (1974a, b) and Reiser and Kobayashi (1974). Fischer's (1971) use of the procedure in analysing alternating priority queues, Gaver's and Shedler's (1973) application in obtaining the processor utilization in a multiprogramming computer system are evidence to the effectiveness of this approximation technique [also see, Fischer (1976, 1977)].

A different approach will be to observe that the process under study converges in some sense to a diffusion process. Iglehart (1965), has shown that in the M/M/n queue if we let the mean inter-arrival time $\to 0$ as $n \to \infty$, then the queue length process (after proper scaling) and normalizing tends to the Ornstein-Uhlenbeck process. McNeil (1973) considers a sequence of non-stationary birth and death processes $\{x_N(t)\}$ with input and output rates dependent on N. He has shown that $\lim_{N \to \infty} x_N(t)$ (after normalizing) corresponds to a non-stationary Ornstein-Uhlenbeck process. An additional reference in this class of efforts is Harrison (1973) who considers a sequence of systems with increasing traffic intensities.

Numerical Approximation:

Numerical approximation can be defined as a simplification which is brought in while actually manipulating the arithmetic expressions leading to an evaluation of certain measure. If we identify an approximation $\hat{x}$ as $\hat{x} = x + \delta$, where x is the corresponding exact

value and $\delta$ is an unknown small quantity, then we call $\hat{x}$ a "point approximation" if $\delta$ is unrestricted in sign, and we call $\hat{x}$ a "one sided approximation" (or an interval approximation) if $\delta$ is restricted in sign. Clearly the more we know about the properties of $\delta$ the more reliable the approximation will be, and we would like $\delta$ to be as small as possible.

The queue G/G/1 presents many difficulties in deriving exact results for its performance measures. Several attempts have been made to obtain approximations. More successful of these are the heavy traffic approximation (a point approximation) and upper and lower bounds (giving an interval approximation) for the mean waiting time given by Kingman [1962, 1965, 1970], Marshall (1968) and Suzuki and Yoshida (1970). All these efforts are based on the fundamental relation

$$W_{n+1} = \max[0, W_n + S_n + T_n]$$

where $W_n$ is the waiting time of the $n^{th}$ customer, $S_n$, is his service time and $T_n$, the time interval between the $(n-1)^{st}$ and the $n^{th}$ customer. Writing $U_n = S_n - T_n$ and denoting the idle period by I, one gets the result

$$E[W] = \frac{E[U^2]}{-2E[U]} - \frac{\pi_0 E[I^2]}{-2E[U]} \tag{1}$$

where $\pi_0$ is the probability that an arrival finds the system empty and $\lim_{n\to\infty} W_n \equiv W$ and $\lim_{n\to\infty} U_n \equiv U$. Since exact values for $E[I]$ and $E[I^2]$ are not available except in cases such as exponential inter-arrival times, an upper bound for $E[W]$ can be obtained as

$$E[W] \leq \frac{\mathrm{Var}[T] + \mathrm{Var}\,[S]}{2(E[T] - E[S])} . \tag{2}$$

When $\rho$ is close to 1, Kingman (1962) has shown that the upper bound for $E[W]$ is a good approximation for $E[W]$ itself. Furthermore, by using central limit theorem on the basic random variables $\{U_n\}$, he

has also shown that under heavy traffic, the waiting time distribution under equilibrium conditions is exponential.

Lower bounds for $E[W]$ have been derived by both Kingman (1970) and Marshall (1968). Marshall shows that

$$E[W] \geq \ell$$

where $\ell$ is the unique solution of the equation

$$x = \int_{-x}^{\infty}[1 - K(u)]du \qquad (x \geq 0)$$

where $P[U \leq u] = K(u)$. Kingman's (1970) alternate bound can be given as

$$E[W] \geq \frac{E[(U^+)^2]}{2(E[T]-E[S])} .$$

where $U^+ = \max\,[0,U]$. Comparing the bounds, Kingman points out that Marshall's bound is sharper in light traffic ($\rho$<<1) whereas his bound is sharper in heavy traffic. Nevertheless, it should be noted that both lower bounds require the knowledge of the distribution of U, whereas the Kingman upper bound depends only on the first two moments of the inter-arrival time and service time distributions. For a concise discussion of bounds and approximations reference can be made to Gross and Harris (1974, ch. 6).

Another approximation for $E[W]$ can be obtained by writing $\pi_0 \simeq 1-\rho$ and $E[I^2] \simeq E[U^2]$, where $\rho$ is the traffic intensity of the system. Then we get from (1),

$$E[W] \simeq \rho\left[\frac{E[T^2] + E[S^2] - 2E[T]E[S]}{2(E[T] - E[S])}\right] \qquad (3)$$

Comparing these approximations for systems with one of the inter-arrival time or service time distributions exponential, Bhat [1974] has shown that the simple approximation given in (3) is in fact better than the heavy traffic approximation given by (2) except when $C_v[S] >> C_v[T]$ where $C_v$ stands for the coefficient of variation.

An additional effort in providing a better approximation for E[W] is that of Marchal (1974), who incorporates the coefficient of variation of the service time distribution $C_v(S)$ by suggesting

$$E[W] \simeq \left(\frac{1+C_v^2[S]}{\rho^{-2}+C_v^2[S]}\right)\left(\frac{Var[T] + Var[S]}{2(E[T] - E[S])}\right)$$

which is identical with the Kingman heavy traffic approximation when $\rho=1$. Marchal has also provided an alternate lower bound

$$E[W] \geq \left[\frac{\rho^2 C_v^2[S] + \rho(\rho-2)}{2(1-\rho)}\right] E[T]$$

which incorporates only $\rho$ and the coefficient of variation of the service time distribution (also see, Marchal [1976, a and b] and Kleinrock [1976, ch. 2]).

Using Martingale theory Ross (1974) has derived upper and lower bounds for the mean delay in the G/G/1 queue. Even though they are somewhat sharper than the ones described above, they are much harder to evaluate.

Extending the Kingman upper bound (2) the following bounds may be given for E[W] in the multi-server queue G/G/s.

$$E[W] \leq \frac{Var[T] + VAR(S/s)}{2(E[T] - E[S/s])}$$

which is essentially the G/G/1 result with a modified service time. This result, originally suggested by Kingman (1965) has been studied by Suzuki and Yoshida (1970). A bound later suggested by Kingman (1970) has the form

$$E[W] \leq \frac{sVar[T] + VAR[S] + (1 - 1/s)\{E[S]\}^2}{2(sE[T] - E[S])}$$

Bounds for some generalizations of the G/G/1 queue have been derived by Marshall (1968b). Some of the cases discussed by him are queues (i) with arrivals in batches of random size, (2) with service in batches of fixed size and (3) with added delay for the first customer in a busy period. Marshall and Wolff (1971) consider bounding the difference between the mean queue length found by an arriving customer and the arbitrary time mean queue length in the G/G/1 system. It is also shown that, for G/G/1, the difference between the mean virtual wait and the mean actual wait does not exceed one half the mean inter-arrival time. Holtzman (1971) derives an upper bound for mean waiting time in a Poisson input single server priority queue by considering waiting time as composed of four distinct parts and obtaining upper bound for each of them.

Heathcote and Winer (1969) take a somewhat different approach in deriving approximations for the moments of waiting times in the G/G/1 queue. Using an expansion related to the central limit theorem, they express $E[W_n] - E[W]$, as an infinite series. Now knowing $E[W]$ one could estimate $E[W_n]$ by approximating the series. Other papers considering approximations and bounds for mean waiting time in G/G/1 and G/G/s queues or their special cases are Granot, and Lemoine (1975), and Harrison (1973).

Approximation techniques have been used for deriving information on other performance measures as well. Rider (1976) approximates the emptiness probability to solve for the average queue size in a time dependent M/M/1 queue. Natvig (1974) approximates the transition probability $P_{10}(t)$ of the transition of the number in the system from 1 to 0 in time t, in a single server Markovian queue with discouragement by simplifying the expression derived through inversion. Beneš (1959) gives an approximation for $p_n$, the probability an arriving

customer finds n busy channels in a G/M/s/s system. In Beneš (1961), an approximation for the covariance function of the number of busy channels in an M/M/s/s system is also provided. Another paper dealing with the approximations for covariance function of the number of busy channels in an M/M/s/s system is Descloux (1965). Approximations for Erlangs loss formula and its derivatives have been given by Jagerman (1974) by truncating a complex series. In these papers, related mostly to teletraffic theory, the technique used is analytical and manipulative. For other papers belonging to this class, readers are referred to Saaty (1961), Cooper (1972), Holtzman (1975) and references cited under Holtzman (1975).

Many of the exact queueing results are given as transform expressions that are difficult to invert. Numerical inversion of Laplace transforms is a convenient technique when such results are needed. Some of the initial papers on this technique are Gaver (1966), Weeks (1966), Dubner and Abate (1968), Chiu, Chen and Huang (1970), and Stehfest (1970). Nance, Bhat and Claybrook (1972) have applied the different methods presented in the above papers to invert the transform of the busy period distribution of an M/G/1/N type queue occurring in a timesharing system. Abate, Dubner and Weinberg (1968) have applied the inversion method to the transform of the waiting time distribution for a mass storage device. It must be pointed out though, that in the process of numerical inversion of transforms it is desirable to experiment with more than one technique since their performance is highly dependent on the original function.

A recent inversion technique given by Knepley and Fischer (1977) discretizes the time parameter and approximates a Laplace Transform by an infinite series. Recursive relations, then provide the needed numerical results. Al-Khayyal and Gross (1977) approximate and bound

the root of the functional equation associated with the GI/M/s queue to give bounds and approximations for stead state measures of effectiveness and probabilities. Another approach based on transforms has been given by Kotiah (1976) for Markovian systems. (These procedures are classified under numerical schemes since the approximation is made on the results of analysis. Nevertheless it is appropriate to mention that the outcome of the procedure is an approximation at the process level.)

Approximation results are also given in the form of limit and convergence theorems. A typical form of a limit theorem is to describe the behavior of a certain process as one of the system parameters approaches a specific limiting value. Convergence in queueing theory has received some attention in the last decade, see for example the survey paper by Iglehart (1973); however, not all such theorems are meant to be used as approximations. Köllerström (1974) shows that the waiting time for the G/G/s system, under some general conditions, converges to a negative exponential as $\rho \to 1$, and then reformulates the result as an approximation with error bounds. Tomko (1972), for $\rho<1$, gives an approximation to the waiting time W(N) in the queue M/M/m/N, in terms of the waiting time W for M/M/m/$\infty$, and provides the rate of convergence. For $\rho=1$, W(N) is shown to converge to a uniform distribution as the capacity $N \to \infty$, and for $\rho>1$, W(N) is shown to converge to a normal distribution. The accuracy for approximating each W(N) with its corresponding asymptotic distribution is also estimated. Kyprianou (1972) shows that the virtual waiting time conditional on its still being in the first busy period in M/G/1 and GI/M/1 is asymptotically, as $\rho \to 1$, gamma distributed with two degrees of freedom and mean 4m, where

$$m = \frac{\mathrm{Var}[T] + \mathrm{Var}[S]}{2(1/E[T] - 1/E[S])} .$$

Schassberger (1970) approximates the G/G/1 queue by a sequence of queues in which the interarrival and service times for the $n^{th}$ system

are Erlangian mixtures, which are convex combinations of Erlangian distributions. He also shows that the distribution function of the virtual waiting time in the $n^{th}$ system converges weakly to that of the original system. Kennedy (1972) proves a similar but more general result for the single server queue.

In a way, the numerical analysis of queueing systems carried out in a series of papers by Neuts (1973) and his co-authors ([Neuts and Klimko (1973 a, b)] can also be identified as an approximating technique. It is a system type of approximation, in that discrete phase type distributions are used for interarrival and service times. In the same spirit one could include papers that have appeared on other numerical aspects of queueing systems such as the solution of Chapman-Kolmogorov equations for birth and death processes. We shall not elaborate on these topics here since the emphasis in this paper is more towards identifying different aspects of approximations.

Validation of Approximations

Validation is an integral part of an approximation procedure. It is needed to support the applicability of the technique and the reliability of results. An applied scientist has to constantly evaluate the trade-off between the ease of application of a particular technique and the accuracy of the ensuing results. Therefore we expect the validation procedure to relate in some way and provide a comparison between approximate and exact results. Generally, validation of approximations can be achieved through error analysis, experimation and simulation. Relative merits of these procedures are discussed in the following paragraphs.

In error analysis, the deviation from an exact result is estimated as a function of the system parameters. For example, if we approximate by truncating a series, any bound on the remainder of this series will bound the error. One of the error analysis procedures is to show that the error converges to zero as one or more of the parameters take a limiting value (see for example Natvig (1974)). If the result is of a limiting nature, then the rate of convergence may help provide an error estimate (Köllerström (1974), Tomko (1972)). For two sided inequality results the error is bounded by the length of the interval; however, one needs to compare the bounds with some exact results as well (Bloomfield and Cox (1972) and Marshall (1968b)). Apart from inequalities, numerical point approximation is the only approach through which error estimates are obtained.

Experimentation is the most common validation technique for approximations. The essential feature of this procedure is to compare the approximated and the exact results for some special cases and if the comparison is favorable similar performance is expected in general. Absence of support for this basis requires careful and exhaustive experimentation covering a wider range of parameters. Clearly this approach can be used for any type of approximation. For example it is used in Beneš (1959), Heathcote and Winer (1969), Holtzman (1971a) and Rider (1976) to validate numerical approximations. Avi-Itzhak and Heyman (1973), Bhat and Fischer (1976), Kuczura (1973), Leibowitz (1961) and Marchal and Harris (1976) have used this method in the context of system approximations. Gaver and Shedler (1973), Heyman (1975), and Reiser and Kobayashi (1974) have used it to validate diffusion approximation technique. The main disadvantage in the procedure, though, is the lack of certainty that the conclusions drawn from experimentation can be extrapolated into more general settings.

Simulating a stochastic system has become popular due to its wide applicability, closeness to reality, and the ability to use statistical analysis techniques. It is the last property that makes simulation a seemingly dependable and appealing validation technique. As can be seen from Fishman (1973) considerable work has been done on the statistical aspects of simulations. But a word of caution is that the analysis is all too often messy and heavily dependent on factors such as sample size. The general approach is to generate samples of the studied process and define estimates for the required measures of performance. If the process is of the regenerative type then the classical statistical techniques can be used to obtain confidence intervals and percentiles (see, for example, Fishman (1974), Iglehart (1975) and Lavenberg and Slutz (1975)). Otherwise, one has to deal with the usual problems arising in simulation such as, dependent samples, effect of initial state, and transient behavior. In either case the simulation model needs validation, and this is usually done through experimentation (see, Rosenshine and Chandra (1975)). Using simulation as a validation technique is common under system approximations (Chandy, Herzog and Woo (1975), Halfin (1975), Moore (1975) and Sauer and Chandy (1975)) and process approximations (Halachmi and Franta (1976), Heyman (1975), Kobayashi (1974a,b) and Reiser and Kobayashi (1974)). For validating numerical approximations even though error analysis is easier, simulation may also be used (see, Descloux (1965)). However, most of these authors have satisfied themselves by the relative size of the percentage difference between the simulated and approximate results. Very few of them have resorted to a statistical analysis of simulated results and provide information such as confidence bounds on their estimates.

While using simulation for validating approximations it is desirable to state the accuracy of the simulated results as well.

As discussed above validation takes different forms that vary in their usefulness. We consider error analysis as the most reliable procedure. However, it is difficult to implement under system and diffusion approximation techniques. Inequalities may not need validation if they are tight enough. Nevertheless, it should be noted that inequalities that are tight are hard to compute and those that are simple to compute are not tight. Thus a sensitivity analysis of inequalities over the rest of the parameters may be recommended (Bhat (1974)). Experimentation and simulation are the more common forms of validation techniques, but while using them, their limitations should be clearly understood.

## Future Prospects

Given above is a broad picture of approximation techniques used in queueing theory. Existing work in the queueing literature has been included in one or the other category of approximation considering the main thrust of the paper. It must be noted, however, that many times, a combination of more than one technique may be needed for a complete solution.

The emergence of approximate results is directly related to the applicability of systems. Furthermore, except for the well-known approximations for the mean waiting time in G/G/1 and related systems, most of the simple and applicable results occur predominantly in application areas of queueing theory such as telephone traffic and computer systems. There is a significant factor in this phenomenon to be noted by a researcher. Since approximate results are obtained for direct use in real world problems, they should be easily computable. Therefore it

does not make sense, except as an intellectual exercise and a theoretical piece of research to provide a better approximation which is much harder to compute, than an available simpler approximation. Thus all applicable approximate results need to be examined from an effort-benefit view point.

As indicated earlier, validation of approximate results has attracted considerable attention, specifically in the application areas. Nevertheless, not enough attention seems to have been paid to the quality of the validation technique itself. In the case of experimentation, more sensitivity analysis is needed. Wider use of statistical techniques related to point and interval estimation should be made when simulation is preferred.

Queueing theory researchers have been criticized for studying systems that are not relavant to the real world. However, it seems to us that this criticism is largely due to the complexity of available results in the literature than due to the systems themselves. If one looks at some of the applied areas of queueing, one finds more complex systems than the general operations research and applied probability literature. The distinction is in the nature of analysis. The results found in the applied area literature are applicable, though approximate. A large percentage of results found in the general literature are less useful though rigorous. Therefore, if we want to keep queueing theory as an integral part of operations research and as a problem solving tool in the general area of applied probability and mathematics, approximation techniques should be put to increasing use whenever necessary. The trend during the past decade is in this direction and there is every reason to believe that this trend is going to continue further in the coming years, bringing more reliability and applicability for the techniques used.

Acknowledgement

The authors are grateful to the referees for pointing out some of the missing references and making constructive comments that have improved the presentation.

REFERENCES

[1] Abate, J., and H. Dubner, and S. B. Weinberg, "Queueing Analysis for the IBM 2314 Disk Storage Facility", J. ACM, 15(4), 577-589, 1968.

[2] Al-Khayyal, F. A., and D. Gross, "On Approximating and Bounding GI/M/c Queues", TIMS Studies in Management Sciences 7, North Holland Pub. Co., 233-245, 1977.

[3] Avi-Itzhak, B., and D. P. Heyman, "Approximate Queueing Model for Multiprogramming Computer Systems", Opns. Res. 21(6), pp. 1212-1230, 1973.

[4] Benes, V. E., "On Trunks With Negative Exponential Holding Times Serving a Renewal Process", Bell Syst. Tech. J., 38, 211-258, 1959.

[5] ______________, "The Covariance Function of a Simple Trunk Group, With Applications to Traffic Measurements", Bell Syst. Tech. J., 40, 117-148, 1961.

[6] Bhat, U. N., "Sensitivity Analysis of Performance Measures in Some Queueing Systems", Tech. Comment No. 30-74, Defense Communication Center, Reston, Virginia, Dec. 1974.

[7] ______________, A. C. Wheeler and M. J. Fischer, "On the Existence of Limiting Distributions in Stochastic Systems with Secondary Inputs", Opsearch, 11 (2, 3), 81-89, 1974.

[8] ______________, and N. U. Prabhu, "A Probabilistic Model for Queues at Traffic Signals", Tech. Rep. IEOR 75007, Dept. of IE/OR, Southern Methodist University, June 1975.

[9] ______________, and M. J. Fischer, "Multichannel Queueing Systems with Heterogeneous Classes of Arrivals", Nav. Res. Log. Qrtl., 23(2), 271-282, 1976.

[10] ______________, and Raju, S., "An Approximate Analysis of a Multi-server Finite Queue with Heterogeneous Customers" Tech. Rep. IEOR 77020, Department of IE/OR, Southern Methodist Univ., Dallas, 1977.

[11] Bloomfield, P., and D. R. Cox, "A Low Traffic Approximation for Queues", J. Appl. Prob. 9, 832-840, 1972.

[12] Brosh, I., "Preemptive Priority Assignment in Multichannel Systems," Opns. Res., 17(3), pp. 526-535, 1969.

[13] Brumelle, S. L., "Some Inequalities for Parallel-Server Queues", Opns. Res., 19 (2), 402-413, 1971.

[14] ______________, "Bounds on the Wait in a GI/M/k Queue", Management Sci., 19(7), 773-777, 1973.

[15] Chandy, K. M., Herzog, N., and L. Woo, "Approximate Analysis of General Queueing Networks", IBM J. Res. Dev., 19(1), 43-49, 1975a.

[16] ____________________________, "Parametric Analysis of Queueing Networks", IBM J. Res. Dev., 19(1), 36-42, 1975b.

[17] Chiu, R. F., Chen, C. F., and C. J. Huang, "A New Method for the Inverse Laplace Transformation Via the Fast Fourier Transform", Twenty-Second Southwestern IEEE Conf. Record, Dallas, Texas, 201-203, 1970.

[18] Conolly, B. ,. "Generalized State Dependent Erlangian Queues: Speculations About Calculating Measures of Effectiveness", J. Appl. Prob., 12, 358-363, 1975.

[19] Cooper, R. B., Introduction to Queueing Theory, MacMillan Co., N. Y., 1972.

[20] Cosmetatos, G. P., "Approximate Equilibrium Results for the Multi-Server Queue (GI/M/n)", Opns, Res. Qurt., 25(4), 625-634, 1974.

[21] Cox, D. R., and H. D. Miller, The Theory of Stochastic Processes, J. W., New York, 1968.

[22] Descloux, A., "On the Accuracy of Loss Estimates", Bell Syst. Tech. J., 44, 1139-1164, July 1965.

[23] Dubner, H., and J. Abate, "Numerical Inversion of the Laplace Transforms by Relating Them to the Finite Fourier Cosine Transform", J. ACM., 15(1), 115-123, 1968.

[24] Feller, W., "An Introduction to Probability Theory and its Applications," Vol. II, J. W., New York, 1966.

[25] Fischer, M. J., "On Two Problems in Alternating Queues", Ph.D., Dissertation, School of Engineering & Applied Science, Southern Methodist Univ., 1971.

[26] ____________, "An Approximation to Queueing Systems with Interruptions," Defense Comm. Eng. Center, Reston, Virginia 22090, 1976.

[27] ________________, "Analysis and Design of Loop Service Systems VIA A Diffusion Approximation", Opns. Res. 25(2), 269-278, 1977.

[28] Fishman, G. S., "Estimation in Multiserver Queueing Simulations", Opns. Res., 22(1), 72-78, 1974.

[29] Fishman, G. S., Concept and Methods in Discrete Event Digital Simulation, John Wiley, New York, 1973.

[30] Gaver, D. P., Jr., "Observing Stochastic Processes and Approximate Transform Inversion", Opns. Res., 14(3), 444-459, 1966.

[31] ______________, "Diffusion Approximations and Models for Certain Congestion Problems," J. Appl. Prob., 5, 607-623, 1968.

[32] ______________, and G. S. Shedler, "Processor Utilization in Multiprogramming Systems Via Diffusion Approximations", Opns. Res., 21(3), 569-576, 1973.

[33] Ghosal, A., Some Aspects of Queueing and Storage Systems, Vol. 23, of Lecture Notes in OR and Math. Systems, Springer Verlag, New York, 1970.

[34] Granot, D., Granot, F., and Lemoine, A., "Approximations for Service System with Nonindependent Interarrival Times", Opns. Res., 23(1), 162-166, 1975.

[35] Gross, D., "Sensitivity of Queueing Models to the Assumption of Exponentiality", Nav. Res. Log. Quart., 22(2), 271-287, 1975.

[36] Gross, D. and C. M. Harris, Fundamentals of Queueing Theory, John Wiley & Sons, N. Y., 1974.

[37] Halachmi, B., and Franta, W. R., "A Diffusion Approximate Solution to the G/K/k Queueing System", Paper presented at the ORSA/TIMS meeting, Nov. 18, 1975.

[38] Halfin, S., "An Approximate Method for Calculating Delays for a Family of Cyclic-Type Queues", Bell Syst. Tech. J., 54(10), 1733-1754, 1975.

[39] Harrison, J. M., "The Heavy Traffic Approximation for Single Server Queues in Series", J. Appl. Prob., 10, 613-639, 1973.

[40] Heathcote, C.R., and Winer, P., "An Approximation for the Moments of Waiting Time," Opns. Res., 17(1), 175-186, 1969.

[41] Heyman, D. P., "An Approximation for the Busy Period of the M/G/1 Queue Using a Diffusion Model", J. Appl. Prob., 11, 159-169, 1974.

[42] _______________, "A Diffusion Model Approximation for the GI/G/1 Queue in Heavy Traffic", Bell Syst. Tech. J., 54(9), 1637-1646, 1975.

[43] Holtzman, J. M., "Bounds for a Dynamic-Priority Queue", Opns. Res., 10(2), 461-468, 1971a.

[44] _______________, "Analysis of Dependence Effects in Telephone Trunking Networks", Bell Syst. Tech. J., 50(8), 2647-2662, 1971b.

[45] _______________, "Some Compatible Approximations in TeleTraffic Theory", Paper presented at the ORSA/TIMS meeting, Nov. 17, 1975.

[46] Iglehart, D., "Limiting Diffusion Approximations for the Many Server Queue and the Repairman Problem", J. Appl. Prob., 2, 429-441, 1965.

[47] ______________, "Weak Convergence in Queueing Theory", Adv. Appl. Prob., 5, 570-594, 1973.

[48] ______________, "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," Nav. Res. Log. Quart., 22, 553-565, 1975.

[49] Jagerman, D. L., "Some Properties of the Erlang Loss Function", Bell Syst. Tech. J., 53(3), 525-551, 1974.

[50] Kennedy, D. P., "The Continuity of the Single Server Queue", J. Appl. Prob., 9, 370-381, 1972.

[51] Kimura, M., "Diffusion Models in Population Genetics", J. Appl. Prob., 1, 177-232, 1964.

[52] Kingman, J. F. C., "Some Inequalities for the Queue GI/G/1", Biometrika, 49, 315-324, 1962.

[53] ________________, "The Heavy Traffic Approximation in The Theory of Queues", Proc. Symp on Congestion Theory (Eds. W. L. Smith and W. E. Wilkinson), Univ. of North Carolina Press, Chapel Hill, 1965.

[54] ________________, "Inequalities in the Theory of Queues", J. Roy. Statist. Soc. B 32, 102-110, 1970.

[55] Kleinrock, L., Queueing Systems Vol. II: Computer Applications, J. W., New York, Chapter 2, 1976.

[56] Knepley, J. E., and Fischer, M. J., "A Numerical Solution for Some Computational Problems Occurring in Queueing Theory," TIMS Studies in Management Sciences 7, North Holland Pub., Co., 271-285, 1977.

[57] Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks I: Equilibrium Queue Distributions", J. ACM, 21(2), 316-328, 1974a.

[58] ________________, "Application of the Diffusion Approximation to Queueing Networks II: Nonequilibrium Distributions and Applications to Computer Modeling", J. ACM, 21(2), 459-469, 1974b.

[59] Kollerstrom, J., "Heavy Traffic Theory for Queues with Several Servers I", J. Appl. Probl, 11, 544-552, 1974.

[60] Kotiah, T. C. T., "Approximations for the Transient Behavior of Some Queues," No. 40, SIU Papers in Math. & Math. Sci., SIU-E, Edwardsville, Ill., 1976.

[61] Kotiah, T. C. T., "On a Linear Programming Technique for the Steady State Behavior of Some Queueing Systems," Opns. Res. 25(2), 289-303, 1977.

[62] Kotiah, T. C. T., J. W. Thompson, and W. A. O'N. Waugh, "Use of Erlangian Distributions for Single-Server Queueing Systems," J. Appl. Prob. 6, 584-593 (1969).

[63] Kuczura, A., "The Interrupted Poisson Process as an Overflow Process", Bell Syst. Tech. J., 52(3), 437-448, 1973.

[64] Kyprianou, E. K., "The Quasi-Stationary Distributions of Queues in Heavy Traffic", J. Appl. Prob., 9, 821-831, 1972.

[65] Lavenberg, S. S., and D. R. Slutz, "Introduction to Regenerative Simulation", IBM J. Res. Develop., 19, 458-462, 1975.

[66] Liebowitz, M. A., "An Approximate Method for Treating a Class of Multi-queue Problems", IBM J. Res. Dev., 5(3), 204-209, 1961.

[67] Luchak, G., "The Solution of the Single Channel Queueing Equation Characterized by a Time Dependent Poisson Distributed Arrival Rate and a General Class of Holding Times," Opns. Res. 4, 711-732, 1956.

[68] Maaloe, E., "Approximation Formulae for Estimation of Waiting-Time in Multiple-Channel Queueing System", Management Sci., 19(6), Appl. Series, 703-710, 1973.

[69] Marchal, W. G., "Some Simple Bounds and Approximations in Queueing", Tech. Mem. T-294, Inst. of Management Sci. & Eng., The George Washington University, Jan., 1974.

[70] ______________, "An Approximate Formula for Waiting Time in Single Server Queues", AIIE Transactions, 8, 473-474, 1976a.

[71] ________________, "Some Simpler Bounds on Mean Queueing Times", Private Communication, 1976b.

[72] ________________, and C. M. Harris, "A Modified Erlang Approach to Approximating GI/G/1 Queues", J. Appl. Prob., 13, 118-126, 1976.

[73] Marshall, K. T., "Some Inequalities in Queueing", Opns. Res., 16(3), 651-665, 1968a.

[74] ________________, "Bounds for Some Generalizations of the GI/G/1 Queue", Oprs. Res, 16(4), 841-848, 1968b.

[75] ________________, and R. W. Wolff, "Customer Average and Time Average Queue lengths and Waiting Times", J. Appl. Prob., 8, 535-542, 1971.

[76] McNeil, D. R., "Diffusion Limits for Congestion Models", J. Appl. Prob., 10, 368-376, 1973.

[77] Moore, S. C. "Approximating The Behavior of Nonstationary Single-Server Queues", Opns. Res., 23(5), 1011-1032, 1975.

[78] Nance, R., Bhat, U. N., and Claybrook, B. G., "Busy Period Analysis of a Time Sharing System: Transform Inversion", J. ACM, 19(3), 453-463, 1972.

[79] Natvig, D., "On the Transient State Probabilities for a Queueing Model Where Potential Customers Are Discouraged by Queue Length", J. Appl. Prob., 11, 345-354, 1974.

[80] Neuts, M. F., "The Single Server Queue in Discrete Time Numerical Analysis Numerical Analysis", Nav. Res. Log. Quart., Part I: 20(2), 297-302, 1973; Part II: (with E. M. Klimko): 20(2), 305-320, 1973; Part III: (with E. M. Klimko): 20(3), 557-568, 1973.

[81] Newell, G. F., "Approximation Methods for Queues with Application to the Fixed Cycle Traffic Light", SIAM Rev., 7(2), 223-239, 1965.

[82] ____________, "Queues with Time-Dependent Arrival Rates I-III," J. Appl. Prob., 5, 436-451, 579-606, 1968.

[83] ____________, Applications of Queueing Theory, Chapman and Hall, London, Chap. 6, 1971

[84] ____________, "Approximate Behavior of Tandem Queues", ITTE Special Report, University of California-Berkeley, 1975.

[85] Nozaki, S. A., and S. M. Ross, "Approximations in Multi-Server Poisson Queues", ORC 76-10, University of California-Berkeley, April 1976.

[86] Reiser, M. and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Qqeueing Systems", IBM J. Res. Dev. 18, 110-124, March 1974.

[87] Rider, K. L., "A Simple Approximation to the Average Queue Size in the Time-Dependent M/M/1 Queue", J. ACM, 23(2), 361-367, 1976.

[88] Rosenshine, M., "Approximating the M/E(k,n)/1 Queue by a Birth-death Process," Paper presented at the ORSA/TIMS meeting, Miami, November, 1976.

[89] Rosenshine, M. and M. J. Chandra, "Approximate Solutions for Some Two State Tandem Queues, Part 1: Individual Arrivals at the Second State", Opns. Res., 23(6), 1155-1166, 1975.

[90] Ross, S. M., "Bounds on the Delay Distribution in GI/G/1 Queues", J. Appl. Prob., 11, 417-421, 1974.

[91] Saaty, T. L., Elements of Queueing Theory, McGraw Hill, N. Y. (1961), p. 159 and p. 162.

[92] Sauer, C. H., and K. M. Chandy, "Approximate Analysis of Central Server Models", IBM J. Res. Dev., 19(3), 301-313, 1975.

[93] Schassberger, R., "On the Waiting Time in the Queueing System GI/G/1", Ann. Math. Statist., 41, 182-187, 1970.

[94] Stehfest, H., "Algorithm 368: Numerical Inversion of Laplace Transforms", Comm. ACM, 13, 47-49, Jan. 1970.

[95] Suzuki, T. and Y. Yoshida, "Inequalities for Many-Server Queue and Other Queues", J. Oper. Res. Soc. of Japan 13, 59-77, 1970.

[96] Tomko, J., "The Rate of Convergence in Limit Theorems for Service Systems with Finite Queue Facility", J. Appl. Prob., 9, 87-102, 1972.

[97] Weeks, W. T., "Numerical Inversion of Laplace Transforms Using Laguerre Functions", J. ACM, 13(3), 419-429, 1966.

[98] Wilkinson, R. I., "Theories For Toll Traffic Engineering in the U.S.A.", Bell Syst. Tech. J., 35(2), 421-514, 1956.

[99] Willemain, T. R., "Approximate Analysis of a Hierarchical Queueing Network", Opns. Res., 22(3), 522-544, 1974.

[100] Wishart, D. M. G., "A Queueing System with Service-Time Distribution of Mixed Chi-Squared Type", Opns. Res. 8, 174-179, 1959.

[101] Yu, O. S., "Stochastic Bounds for Heterogeneous-Server Queues with Erlang Service Times", J. App. Prob., 11, 785-796, 1974.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: IEOR-76017 | 2. GOVT ACCESSION NO. |
| | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) 6 Approximation Techniques in the Solution of Queueing Problems | 5. TYPE OF REPORT & PERIOD COVERED 9 Technical Rep[redacted]t, |
| | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) 10 U. Narayan Bhat, Martin J. Fischer, Mohamed Shalaby | 8. CONTRACT OR GRANT NUMBER(s) 15 N00014-75-C-0517 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University, Dallas, Texas 75275 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR042-324-436 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Statistics and Probability Program, Office of Naval Research, Department of the Navy, Arlington, Va. 22217 | 12. REPORT DATE 11 October 1976, Mar[redacted]78 |
| | 13. NUMBER OF PAGES 32 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | 15. SECURITY CLASS. (of this report) 12 Unclassified 36 p. |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale. It s distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

System approximation
Process approximation
Numerical approximation
Validation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In the study of complex queueing systems analysis techniques aimed at providing exact solutions become ineffective. Approximation techniques provide an attractive alternative in such cases. This paper gives an overview of different types of approximation techniques available in the literature and points out their relative merits. Also the need for proper validation procedures of approximation techniques is emphasized.

409 434

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)